Multilevel Sequential Monte Carlo with Dimension-Independent Likelihood-Informed Proposals

DOI:
10.1137/17M1120993

Document Version
Final published version

SIAM/ASA J. UNCERTAINTY QUANTIFICATION
Vol. 6, No. 2, pp. 762–786

Multilevel Sequential Monte Carlo with Dimension-Independent Likelihood-Informed Proposals*

Alexandros Beskos†, Ajay Jasra‡, Kody Law§, Youssef Marzouk¶, and Yan Zhou‡

Abstract. In this article we develop a new sequential Monte Carlo method for multilevel Monte Carlo estimation. In particular, the method can be used to estimate expectations with respect to a target probability distribution over an infinite-dimensional and noncompact space—as produced, for example, by a Bayesian inverse problem with a Gaussian random field prior. Under suitable assumptions the MLSMC method has the optimal $\mathcal{O}(\varepsilon^{-2})$ bound on the cost to obtain a mean-square error of $\mathcal{O}(\varepsilon^2)$. The algorithm is accelerated by dimension-independent likelihood-informed proposals [T. Cui, K. J. Law, and Y. M. Marzouk, (2016), *J. Comput. Phys.*, 304, pp. 109–137] designed for Gaussian priors, leveraging a novel variation which uses empirical covariance information in lieu of Hessian information, hence eliminating the requirement for gradient evaluations. The efficiency of the algorithm is illustrated on two examples: (i) inversion of noisy pressure measurements in a PDE model of Darcy flow to recover the posterior distribution of the permeability field and (ii) inversion of noisy measurements of the solution of an SDE to recover the posterior path measure.

Key words. multilevel Monte Carlo, sequential Monte Carlo, Bayesian inverse problem, uncertainty quantification

AMS subject classifications. 82C80, 60K35

DOI. 10.1137/17M1120993

1. Introduction.

The estimation of expectations under a target probability distribution over an infinite-dimensional and noncompact space has a wide range of applications, e.g., [33] and the references therein. In particular, Bayesian inverse problems (BIP) with Gaussian random field priors are an important class of such mathematical models. In most cases of practical interest, one must compute estimates of expectations using the Monte Carlo method

*Received by the editors March 14, 2017; accepted for publication (in revised form) February 5, 2018; published electronically June 5, 2018.

http://www.siam.org/journals/juq/6-2/M112099.html

Funding: The work of the first author was supported by the Leverhulme Trust Prize. The work of the second and fifth authors was supported by Ministry of Education AcRF tier 2 grant, R-155-000-161-112. The work of the second author was also supported under the KAUST Competitive Research Grants Program-Round 4 (CRG4) project, Advanced Multi-Level sampling techniques for Bayesian Inverse Problems with applications to subsurface, ref: 2584. The work of the third author was supported by an ORNL LDRD Strategic Hire and also in part by the US Department of Energy (DOE) Office of Advanced Scientific Computing Research under grant DE-SC0009297 (DiaMond MMICC). The work of the fourth author was supported in part by the US Department of Energy (DOE) Office of Advanced Scientific Computing Research under grant DE-SC0009297 (DiaMond MMICC).

†Department of Statistical Science, UCL, London, UK (a.beskos@ucl.ac.uk).

‡Department of Statistics & Applied Probability, NUS, 117546, Singapore (staja@nus.edu.sg, stazhou@nus.edu.sg).

§School of Mathematics, University of Manchester, Manchester, UK (kodylaw@gmail.com).

¶Department of Aeronautics & Astronautics, MIT, Cambridge, MA 02139 (ymarz@mit.edu).

under a finite-dimensional discretization of the associated probability distribution; see [8, 23], for example.

In many scenarios, such as the BIP above, the finite-dimensional approximation of the probability distribution of interest becomes more accurate but more computationally expensive as the dimension of the approximation increases to infinity. This is precisely the class of problems of interest in this paper. It is well known that the multilevel Monte Carlo (MLMC) method [17, 21] can reduce the computational effort, relative to independent sampling (from a given fixed resolution) required to obtain a particular mean-square error; see [23, 4]. MLMC uses a sequence of increasingly accurate approximations of the target distribution and relies on sampling *independently* from a collection of couples of this sequence and employing the multilevel (ML) identity; details are given later in the paper. A main challenge in the problems of interest here is that such independent sampling is not currently possible.

This paper employs sequential Monte Carlo (SMC) samplers, as these approaches have been shown to outperform Markov chain Monte Carlo (MCMC) in many cases (e.g., [25]) and to be robust in classes of high-dimensional problems [1, 2, 13]. In [4] an SMC method for ML estimation was introduced (and extended in [14]) and analyzed for a class of BIPs. This method was developed for scenarios where ML estimation is expected to be quite beneficial, but where independent sampling from the couplings of interest is not trivial to perform. The papers [4, 14] use SMC to replace independent sampling and coupling in the ML context. However, the approaches in [4, 14] can only deal with a sequence of probability distributions on a *fixed* state-space. That is, the dimension of the parameter of interest, and hence the state space of the resulting sequence of distributions, is assumed to be fixed. Different levels in the estimation scheme correspond to refinements of the PDE approximation for the forward model. In contrast, this paper assumes that the parameter of interest is in principle infinite-dimensional; thus the resolution of the parameter is refined along with the approximation of the PDE model as the level increases. The dimension of the state space of the resulting distributions therefore increases at each level, and hence a modification of previous multilevel algorithms in [4, 14] is required.

The main contributions of this paper are as follows:

1. the design of a new SMC sampler approach for MLMC estimation which allows refinement of the parameter space and solver of the forward model,
2. under assumptions, a theoretical cost analysis for this MLSMC method,
3. introduction of a covariance-based version of the likelihood-informed subspace (cLIS) of [10, 9] and a method for its sample approximation,
4. adoption of efficient dimension-independent likelihood-informed (DILI) proposals [9] within the SMC algorithm, utilizing the new cLIS.

In terms of 1 and 2 there are few methodologies that exist to implement the ML procedure on our problem of interest, nor accompanying theory. For 3 the idea of the approach is to improve proposals in MCMC. In comparison to the LIS method of [10], cLIS does not need potentially expensive or complex gradient/Hessian information. In addition, as we discuss in section 3, when used as part of Gaussian proposals, it is not clear that there is a considerable loss in information of using cLIS against LIS. SMC samplers rely on MCMC as well as on sequential importance sampling/resampling. For such samplers to work well, the MCMC step

must mix over the high-dimensional state space at a reasonable rate. We show that this can be achieved through the mechanism in point 4.

A related multilevel approach in [19] could also be considered for the class of problems in this paper. However, it would have to be modified for the present setting. The sequence of distributions considered in this paper is associated with a convergent approximation of a single *static* target distribution. The present work is not relevant for *online* data assimilation.

This article is structured as follows. In section 2, the basic algorithm and estimation procedure are introduced. Section 3 presents the derivation of the DILI proposal given a collection of samples. Section 4 shows how the DILI proposal methodology can be embedded within the context of MLSMC. Section 5 presents several numerical implementations of our methodology. Some proofs and technical mathematical results are deferred to the appendix.

2. Multilevel sequential Monte Carlo samplers.

2.1. Model. Let $U_0, U_1, \ldots$ be a sequence of spaces, $U_n \subseteq \mathbb{R}^{d'_n}$, $d'_n \in \mathbb{N}$, $n \geq 0$. Let $E_n = \bigotimes_{i=0}^{n} U_i \subseteq \mathbb{R}^{d_n}$, where $d_n = \sum_{i=0}^{n} d'_i$. We consider a sequence of probability measures $\{\hat{\eta}_n\}_{n\geq 0}$ on $\{E_n\}_{n\geq 0}$. Without confusion, we denote the densities with respect to appropriate dominating measures (for this work, these will correspond to Lebesgue measures) also as $\{\hat{\eta}_n\}_{n\geq 0}$. We suppose that

$$\hat{\eta}_n(u_{0:n}) = \frac{\kappa_n(u_{0:n})}{Z_n}$$

with $\kappa_n : E_n \to \mathbb{R}^+$ known but Z_n possibly unknown. The probability measures of interest in this work are associated with Bayesian problems in high-dimensions. (BIPs over a basis function-type approximate solution of a PDE or inference problems related with the approximation of an SDE; see section 5.) As n grows, so does the dimension of the target, towards a well defined infinite-dimensional limit. Let the approximate forward solution of the continuous system associated to an input $u_{0:\ell} \in E_\ell$ processed into a finite number $p \in \mathbb{N}$ of summary values be denoted by ρ_ℓ, i.e., $\rho_\ell : E_\ell \to \mathbb{R}^p$. We are interested in computing, for bounded-measurable functions $\varphi : \mathbb{R}^p \to \mathbb{R}$,

$$\hat{\eta}_L(\varphi \circ \rho_L) := \int_{E_L} \varphi(\rho_L(u_{0:L}))\hat{\eta}_L(u_{0:L})du_{0:L}$$

for some large L or, ideally, $\hat{\eta}_\infty(\varphi \circ \rho_\infty)$. We denote the infinite resolution expectation as $\hat{\eta}(\varphi) := \hat{\eta}_\infty(\varphi \circ \rho_\infty)$. In addition it is of interest to estimate the normalizing constant Z_L or Z_∞. Define $\rho_l(u_{0:n}) := \rho_l(u_{0:l})$ for $n > l$.

Assume that

$$\kappa_\ell(du_{0:\ell}) = \mathcal{L}_\ell(u_{0:\ell})\mu_0(du_{0:\ell}), \tag{2.1}$$

where $\mathcal{L}_\ell(u_{0:\ell})$ is a likelihood term, related to observations of the approximate solution of the continuous system, and μ_0 is the prior density typically defined on the whole E_∞ with $\mu_0(du_{0:\ell})$ referring to its finite-dimensional marginal distribution on $u_{0:\ell} \in E_\ell$. It is worth noting that the algorithms to be described later will be more broadly applicable than the context described in this paragraph.

2.2. Algorithm. We consider a sequence of Markov kernels $\{K_n\}_{n\geq 0}$, $K_n : E_n \to \mathcal{P}(E_n)$ ($\mathcal{P}(E_n)$ is the set the probability measures on E_n) each keeping the respective measure $\{\hat{\eta}_n\}_{n\geq 0}$ invariant, i.e., $\hat{\eta}_n K_n = \hat{\eta}_n$. Let $\{q_n\}_{n\geq 1}$ be a sequence of probability kernels on $\{U_n\}_{n\geq 1}$, so that $q_n : E_{n-1} \to \mathcal{P}(U_n)$. Let $\{M_n\}_{n\geq 1}$, $M_n : E_{n-1} \to \mathcal{P}(E_n)$ be defined as

$$M_n(u_{0:n-1}, du'_{0:n}) = K_{n-1}(u_{0:n-1}, du'_{0:n-1}) \otimes q_n(u'_{0:n-1}, du'_n).$$

Finally, let $G_0(u_0) = 1$, and for $n \geq 1$

$$G_n(u_{0:n}) = \frac{\kappa_n(u_{0:n})}{\kappa_{n-1}(u_{0:n-1}) q_n(u_{0:n-1}, u_n)},$$

where the slightly degenerate notation $q_n(u_{0:n-1}, du_n) = q_n(u_{0:n-1}, u_n) du_n$ has been used. For $n \geq 0$, $\varphi \in \mathcal{B}_b(E_n)$ (the space of bounded, measurable functions on E_n), set

$$\gamma_n(\varphi) := \int_{E_0 \times \cdots \times E_n} \varphi(u_{0:n}(n)) \Big\{ \prod_{p=0}^{n-1} G_p(u_{0:p}(p)) \Big\} \hat{\eta}_0(du_0(0)) \Big\{ \prod_{p=1}^{n} M_p(u_{0:p-1}(p-1), du_{0:p}(p)) \Big\}.$$

Then, from standard Feynman–Kac model calculations [11] one can show that, for $n \geq 1$, $\hat{\eta}_n(\varphi) = \gamma_n(G_n\varphi)/\gamma_n(G_n)$. Denote $\eta_n(\varphi) = \gamma_n(\varphi)/\gamma_n(1)$, $n \geq 0$. Note that $Z_n/Z_0 \equiv \gamma_n(G_n)$ and $\eta_0 \equiv \hat{\eta}_0$ (since $G_0 \equiv 1$). Let $n \geq 1$, $\mu \in \mathcal{P}(E_{n-1})$ and define $\Phi_n : \mathcal{P}(E_{n-1}) \to \mathcal{P}(E_n)$

$$\Phi_n(\mu)(du_{0:n}) = \frac{\mu(G_{n-1} M_n(\,\cdot\,, du_{0:n}))}{\mu(G_{n-1})}.$$

Our ML algorithm works as follows. Let $N_0 \geq N_1 \geq \cdots \geq N_L \geq 1$ be a sequence of given integers. The algorithm approximates the sequence $\{\eta_n\}_{0\leq n\leq L}$. At time zero, one samples

$$\prod_{i=1}^{N_0} \eta_0(du_0^i(0)).$$

Let $\eta_0^{N_0}$ denote the N_0-empirical measure of samples. At time 1, one samples from

$$\prod_{i=1}^{N_1} \Phi_1(\eta_0^{N_0})(du_{0:1}^i(1)).$$

Thus, in an obvious extension of the notation, the joint law of the algorithm is

$$\Big(\prod_{i=1}^{N_0} \eta_0(du_0^i(0)) \Big) \Big(\prod_{\ell=1}^{L} \prod_{i=1}^{N_l} \Phi_\ell(\eta_{\ell-1}^{N_{\ell-1}})(du_{0:\ell}^i(\ell)) \Big).$$

Notice that the present algorithm is different from the one in [12], and hence also the algorithm in [4]. In particular, the state space dimension here grows at each iteration. Moreover, the present algorithm is *not* the standard one used for nonlinear state-space models, as our implementation includes an MCMC step and as we now explain, this is not desirable in these

alternative contexts. The present algorithm will have increasing cost with time (the subscript n). This is because the cost of the MCMC steps and sometimes the cost of computing G_n will grow at some rate with the size of the state space. This is generally not desirable for classical applications of SMC methods associated with the filtering of non-Gaussian and nonlinear state-space models (i.e., dynamic problems with data arriving sequentially in time). The algorithm described above is designed for inverse problems that are so-called static, i.e., one has a single instance of the data from which to make inference. Such growth is therefore less of a concern. Note, also, that the MCMC step is *necessary* for the efficiency of the algorithm. This algorithm can be iterated until $n = \infty$, although the cost will be infinite. Here we are interested in controlling the rate of growth in cost as one approaches the unbiased estimator. In certain contexts it is possible to design an algorithm which produces unbiased estimators targeting E_∞ for finite cost; see [2]. The extension of this algorithm to the latter context is beyond the scope of the current article and will be a topic of future work.

2.2.1. Multilevel estimation. We note that

$$\hat{\eta}_L(\varphi \circ \rho_L) = \sum_{\ell=0}^{L} [\hat{\eta}_\ell(\varphi \circ \rho_\ell) - \hat{\eta}_{\ell-1}(\varphi \circ \rho_{\ell-1})]$$

with $\hat{\eta}_{-1}(\varphi \circ \rho_{-1}) := 0$. Now

$$\hat{\eta}_\ell(\varphi \circ \rho_\ell) = \frac{Z_{\ell-1}}{Z_\ell} \hat{\eta}_{\ell-1} M_\ell(G_\ell\, \varphi \circ \rho_\ell) = \frac{Z_{\ell-1}}{Z_\ell} \left(\hat{\eta}_{\ell-1} \otimes q_\ell\right)(G_\ell\, \varphi \circ \rho_\ell).$$

Also $\eta_\ell(\varphi \circ \rho_{\ell-1}) \equiv \hat{\eta}_{\ell-1}(\varphi \circ \rho_{\ell-1})$. So one can approximate, for $\ell \geq 1$, $\hat{\eta}_\ell(\varphi \circ \rho_\ell) - \hat{\eta}_{\ell-1}(\varphi \circ \rho_{\ell-1})$ by

$$\eta_\ell^{N_\ell}(G_\ell)^{-1} \eta_\ell^{N_\ell}(G_\ell\, \varphi \circ \rho_\ell) - \eta_\ell^{N_\ell}(\varphi \circ \rho_{\ell-1}),$$

and $\hat{\eta}_0(\varphi \circ \rho_0)$ by $\eta_0^{N_0}(\varphi \circ \rho_0)$. This estimate is *different* than that in [4], but similar in spirit. As $Z_\ell / Z_0 = \gamma_\ell(G_\ell)$, this can be approximated by

$$\gamma_\ell^{N_{0:\ell}}(G_\ell) = \prod_{l=0}^{\ell} \eta_l^{N_l}(G_l).$$

As shown in [14] (in a different context) this estimator has similar properties to one that follows the "standard" ML type principle.

Define

$$\hat{\eta}_L^{\mathrm{ML}}(\varphi) := \eta_0^{N_0}(\varphi \circ \rho_0) + \sum_{\ell=0}^{L} \eta_\ell^{N_\ell}(G_\ell)^{-1} \eta_\ell^{N_\ell}(G_\ell\, \varphi \circ \rho_\ell) - \eta_\ell^{N_\ell}(\varphi \circ \rho_{\ell-1}).$$

Let $a(\epsilon) \lesssim b(\epsilon)$ denote (for nonnegative $a(\epsilon)$, $b(\epsilon)$) that there exists a constant $c > 0$ such that $a(\epsilon) \leq c\, b(\epsilon)$ for all ϵ sufficiently small. The following proposition is proved in the appendix.

Proposition 2.1. *Under appropriate assumptions (see assumptions* A1–A3 *in the appendix), for any $\varepsilon > 0$ there exists an L and sequence $\{N_\ell\}_{\ell=0}^{L}$, such that*

$$\mathbb{E}|\hat{\eta}_L^{\mathrm{ML}}(\varphi) - \hat{\eta}(\varphi)|^2 \lesssim \varepsilon^2, \tag{2.2}$$

for a total computational cost $\mathrm{Cost} \lesssim \varepsilon^{-2}$.

3. Dimension-independent likelihood-informed proposals. In this section, we describe how to set up DILI proposals [9] that will later on be embedded into the MLSMC framework. In particular, section 3.1 describes the nonintrusive (i.e., gradient-free) covariance-based construction of cLIS. Section 3.2 illustrates an approach for obtaining the dimension of cLIS. Section 3.3 provides a description of the resulting DILI proposals which will ultimately be used as the kernels K_ℓ within the MLSMC algorithm. Later on, section 4 will place the cLIS construction within the ML context. In section 5, we will present simulation studies that illustrate the significance of such a likelihood-informed proposal for the effectiveness of the overall MLSMC method.

3.1. Sample approximation of the cLIS. For simplicity of exposition, this section will consider some fixed (high) finite dimension $d \geq 1$; however the framework is easily extended to infinite-dimensional spaces. Consider the case where we have a particle population $u^i \in \mathbb{R}^d$, $1 \leq i \leq N$, for some $N \geq 1$, from a probability measure $\nu \in \mathcal{P}(\mathbb{R}^d)$. Define the covariance matrix

$$C := \mathbb{E}_\nu[(u - \mathbb{E}_\nu(u)) \otimes (u - \mathbb{E}_\nu(u))],$$

and assume that

$$C = I - QQ^\top. \tag{3.1}$$

Here $Q \in \mathbb{R}^{d\times m}$, I is the $d \times d$ identity matrix, and $m \ll d$ is the dimension of a linear subspace of *concentration* of the measure ν with respect to a reference measure ν_0, where the latter has identity covariance,

$$\mathbb{E}_{\nu_0}[(u - \mathbb{E}_{\nu_0}(u)) \otimes (u - \mathbb{E}_{\nu_0}(u))] = I.$$

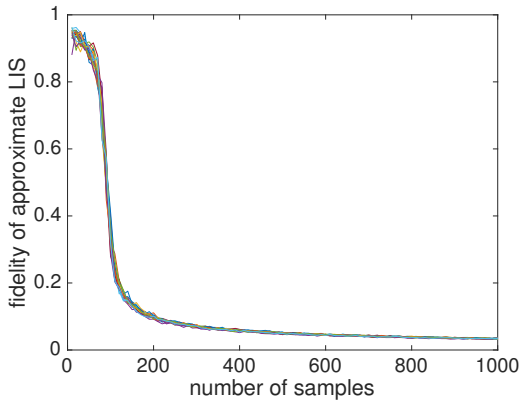
One should think of ν_0 and ν as prior and posterior measures, respectively, in a given context. The column space of Q, defined by the matrix P of m orthonormal eigenvectors of Q such that $Q = P\Lambda^{1/2}$ for some full rank diagonal matrix of positive eigenvalues $\Lambda \in \mathbb{R}^{m\times m}$, is a covariance-based generalization of the gradient-based LIS introduced in [10, 9] and will be referred to below as a cLIS. Notice that the condition above is equivalent to

$$C^{-1} = I + MM^\top,$$

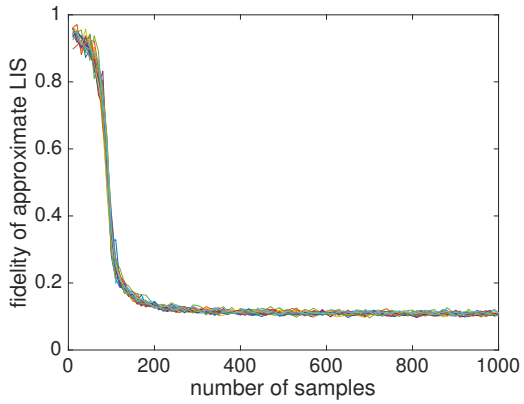
where M and Q have the same column space. If $\sigma^2_{Q,i}$ and $\sigma^2_{M,i}$ are the squared singular values of the matrices Q and M, respectively, then $\sigma^2_{M,i} = \sigma^2_{Q,i}/(1 - \sigma^2_{Q,i})$ for $1 \leq i \leq m$, once the appropriate ordering is applied [32].

In practical settings, (3.1) may hold only approximately, in the sense that $C \approx I - QQ^\top$ (see [9] for applications); however for simplicity of presentation, here we will assume that it holds exactly. The case where (3.1) does not hold is beyond the scope of the present work.

We want to estimate C and, more importantly, the column space of Q, using the particles $\{u^i\}_{i=1}^N$. The simplest way this can be done is the following. Assume for simplicity that we know the rank m. We construct a sample approximation of the low-rank correction to the



(a) cLIS fidelity as a function of the number of samples for 20 random targets, using i.i.d. samples.



(b) cLIS fidelity as a function of the number of samples for random targets, using correlated samples with IACT=100.

Figure 1.

covariance as

$$H_N := I - \frac{1}{N-1}\sum_{i=1}^{N}(u^i - \bar{u}) \otimes (u^i - \bar{u}), \qquad \bar{u} = \frac{1}{N}\sum_{i=1}^{N} u^i. \tag{3.2}$$

Now, we use an iterative algorithm, such as the Lanczos iteration, to compute the dominant m eigenpairs giving rise to $P_{N,m} \in \mathbb{R}^{d\times m}$, and a diagonal $\Lambda_{N,m} \in \mathbb{R}^{m\times m}$ (with the diagonal comprised of the m dominant eigenvalues) so that $H_N \approx P_{N,m}\Lambda_{N,m}(P_{N,m})^\top$. The (orthonormal) columns of $P_{N,m}$ correspond to the N-sample approximation of the m-dimensional cLIS. Simulations indicate that as long as $N > d$, this approach provides a reasonable approximation of the cLIS. Indeed, (3.1) may be seen as an inverse version of the *spiked covariance model* from [15]. There it was shown that this is in fact the required number of samples as $d \to \infty$, and explicit error bounds are provided. See also [28] for further exploration of this point.

For a simple example, see Figure 1, where we consider 20 random Gaussian targets with $d = 100$ of known rank $m = 10$, i.e., a 10-dimensional (IOD) cLIS. The random Gaussian targets were constructed as follows. For $k = 1, \ldots, 20$, $i = 1, \ldots, d$, $j = 1, \ldots, m$, let $A_{ij}^{(k)} \sim N(0, 1)$, independently over i, j. Let $C^{(k)} = (A^{(k)}A^{(k),\top} + I_d)^{-1}$. The kth Gaussian is given by $N(0, C^{(k)})$, $k = 1, \ldots, 20$. The cLIS is approximated using (3.2) with independent and identically distributed (i.i.d.) samples (Figure 1, left panel) and highly correlated samples (Figure 1, right panel), and the cLIS fidelity is approximated using

$$\text{fidelity} := \|PP^\top(I - P_{N,m}P_{N,m}^\top)\|/\|PP^\top\|, \tag{3.3}$$

where P is the matrix with orthonormal columns making up the exact m-dimensional cLIS (analytically known in this synthetic example) and $\|\cdot\|$ indicates the Frobenius norm. The rationale behind this nonsymmetric subspace divergence is that we are particularly concerned with how well $P_{N,m}$ approximates P, i.e., with the projection of $P_{N,m}$ onto P. Note that a

weighted subspace distance [27] as advocated in [10] can be used to favor recovery of the most important directions of the cLIS; alternatively one might use a modification of the Förstner [16] metric between SPD matrices, as proposed in [9]. Also, note that these ideal error metrics cannot be computed in practice, since we do not have access to P.

Remark 3.1. In general, the cLIS construction may miss local features that can be captured by the original gradient-based LIS of [10]. However, cLIS will ultimately be used here only for the construction of a *Gaussian* proposal, and it is unclear what benefit a more sensitive gradient-based LIS would offer for this purpose. For example, consider a d-dimensional distribution that is bimodal along a data-informed direction with unnormalized density $\exp\{-((y_1-u_1^2)^2+(y_2-u_2)^2)/(2\gamma^2)-\frac{1}{2}|u|^2\}$. As $\gamma \to 0$, the averaged Hessian used to build an LIS in [9] will be large across (u_1, u_2), which both are clearly informed by the data. The cLIS, though, will only identify u_2. Nonetheless, a global Gaussian proposal constructed using *either* subspace will have difficulty sampling the target.

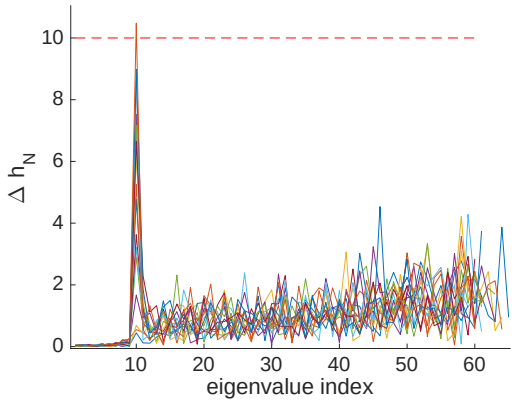
3.2. Estimating the dimension of the cLIS. It is critical to develop a method to automatically estimate the cLIS dimension m in realistic scenarios, where one may know or suspect that there exists a low-dimensional subspace informed by the data of some unknown dimension $m \geq 1$. For this task, the following algorithm is proposed.

Let $\tilde{h}_N$ denote the full vector of d eigenvalues of matrix H_N in (3.2), sorted in decreasing order, and let $h_N = \tilde{h}_N \mathbf{1}_{\{\tilde{h}_N \geq 0\}}$. We truncate the negative eigenvalues, as there may be some large negative eigenvalues when the sample covariance approximation is poor, while the cLIS approximation can actually already be adequate. This also prevents issues arising when a perturbation from the prior is not negative definite, as might occur with multimodal posteriors. Now define, for $i = 1, \dots d-1$,

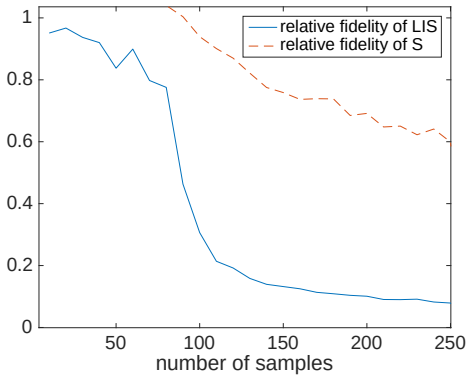
$$(\widetilde{\Delta h_N})_i = |h_{N,i+1} - h_{N,i}|, \quad \overline{\Delta h_N} = \frac{1}{d-1}\sum_{i=1}^{d-1}(\widetilde{\Delta h_N})_i, \quad (\Delta h_N)_i = (\widetilde{\Delta h_N})_i / \overline{\Delta h_N}.$$

It will suffice to find the index i_{ex} such that $(\Delta h_N)_{i_{\text{ex}}} > \text{TOL}$, where TOL is some prespecified reasonable value in between the sample error and the expected size of the gap in the spectrum at convergence. This index, effectively the position where the relative absolute difference in the eigenvalues delivers a "spike," is then taken as the estimate of m, i.e., $\hat{m} = i_{\text{ex}}$.

Figure 2 applies this approach to the target $N(0, C^{(1)})$, where $C^{(1)}$ is constructed as described above, i.e., one of the targets from Figure 1. The left panel of Figure 2 illustrates the growth of the gap in spectrum beyond the threshold for the target $N(0, C^{(1)})$ from Figure 1, where the horizontal axis indicates the index of the nonzero values of h_N, and the connected lines of different color show the values of vector Δh_N for different choices of sample size $N = d, d+m, \dots, 250$. The threshold value is set to TOL $= 10$, and $(\Delta h_N)_{i_{\text{ex}}}$ exceeds this value already for $N = 250$ samples with the correct value of $m = i_{\text{ex}} = 10$. It is clear that in this case the increments Δh_N show a spike at the correct value $m = 10$ for a large enough sample size. Notice also that the right panel in Figure 2 illustrates that, in this example, accurate-enough sample approximation of the cLIS is less of a challenge than sample approximation of the covariance matrix.



(a) Sequence of increments in vector Δh_N for various choices of sample size N, for a given target distribution $N(0, C^{(1)})$. The colors are arbitrary. The point is that, as N increases, the noise is reduced and the peak sharpens.



(b) Relative fidelity of the cLIS approximation, i.e., (3.3) (blue), in comparison to the relative fidelity of the sample covariance S, i.e., $\|S - C\|/\|C\|$ (red dashed) in terms of the number of samples.

Figure 2.

3.3. Use of a subspace at a mutation step. Mutation steps in our SMC algorithm will use a DILI proposal, defined abstractly as follows. Consider a subspace determined by the collection of orthonormal vectors $P = [e_1, e_2, \ldots, e_m]$, spanning an m-dimensional subspace of $\mathbb{R}^d$, together with an approximation of the projected mean $\bar{u} \approx PP^\top \mathbb{E}_\eta u$ and the covariance of the coordinates $(\langle u, e_i \rangle)_{i=1}^m$, $\Sigma \approx P^\top C P \in \mathbb{R}^{m \times m}$. We will make use of the orthogonal decomposition

$$u = PP^\top u + (I - PP^\top)u,$$

where $PP^\top u$ is the orthogonal projection of u onto the subspace. Let $u' \sim Q(u, \cdot)$ be defined by

$$u' = \bar{u} + A(u - \bar{u}) + Bw, \quad w \sim N(0, I), \tag{3.4}$$

where we have defined

$$A = P(I_m - b_m \Sigma)^{1/2} P^\top + (1 - b_\perp^2)^{1/2}(I - PP^\top), \tag{3.5}$$

$$B = P\sqrt{b_m \Sigma} P^\top + b_\perp (I - PP^\top), \tag{3.6}$$

and $b_m, b_\perp \in (0, 1)$ are small step sizes on and off the subspace, respectively. The second summands in (3.5)–(3.6) correspond to a preconditioned Crank–Nicolson (pCN) step on the space orthogonal to the subspace, while the first summands correspond to a step that uses the covariance Σ to scale the step sizes across the various directions of the subspace. All matrix operations are carried out via the eigendecomposition of the symmetric, positive semidefinite Σ. The matrices weighting the proposal satisfy $A^2 + B^2 = I$ and take into account appropriately the covariance (likelihood) information. The proposal $Q(u, \cdot)$ is reversible with respect

to $\nu_0 \equiv N(0, I)$, in the sense that

$$\nu_0(du)Q(u, dv) = \nu_0(dv)Q(v, du).$$

(Note this implies $\int_E \nu_0(du)Q(u, dv) = \nu_0(dv)$.) The above proposal therefore provides an effective *dimension-independent* (DI) proposal for the whitened Gaussian prior μ_0, as the algorithm is well-defined even on infinite-dimensional separable Hilbert spaces (i.e., even if $d = \infty$). In the case of nonwhite priors, e.g., for the standard assumption of a covariance that is a trace-class operator, one must simply employ a change of variables. (See [9] for more details on this construction.)

Recall the assumption from section 3.1 that the covariance is a negative semidefinite perturbation of the prior, so that if Σ is the exact covariance, then $I_m - b_m\Sigma$ is guaranteed to be positive semidefinite (and vanishing only off the true cLIS and when $b_m = 1$). When the approximate cLIS is constructed from samples in practice, one must take care to ensure the nonnegativity of $I_m - b_m\Sigma$.

Note that as long as the proposal is split according to a rotation induced by an operator P with a finite range m, then *any proposal* can be used on the subspace spanned by P and the DI property will be preserved. However, the proposal should be chosen such that the resulting Metropolis–Hastings algorithm is convergent, as the above algorithm is proved to be in [31]. If derivatives were available, we could use them on the cLIS part of the proposal above to construct manifold-based proposals, as was recently done in [3]. The following proposal, which preserves the Gaussian approximation of the posterior on the cLIS (instead of the prior) is not, in general, geometrically ergodic

$$A = (1 - b_m)^{1/2}PP^\top + (1 - b_\perp^2)^{1/2}(I - PP^\top),$$
$$B = P\sqrt{b_m\Sigma}P^\top + b_\perp(I - PP^\top).$$

In particular, it is shown in [29] Theorem 2.1 that this proposal is not ergodic for $b = b_m = b_\perp = 1$, for a wide range of target distributions, including Gaussians with a covariance larger than Σ on the subspace. This property is expected to hold for $b_m < 1$ as well. In [6] it is suggested simply to scale the covariance Σ by a factor $(1 + \epsilon)$ for $\epsilon > 0$. This strategy works in practice, but the downside is that there is no clear criterion for the choice of ϵ.

4. Multilevel cLIS in MLSMC samplers.

We will now embed the cLIS methodology within a multilevel sampling framework. The idea here is that the cLIS is expected to converge at some level of mesh refinement that is less accurate than the final level required by the MLSMC algorithm, so that the cLIS can then be embedded into higher levels at a nominal cost. See [10] for an example of the LIS basis converging under mesh refinement. Furthermore, the telescopic identity can be leveraged along the way to improve the cost of the algorithm. It is noted that cLIS-based proposals can be used outside of the MLSMC context as well, and this will be the subject of future work (for instance, as part of standard SMC samplers).

Section 4 will be organized as follows. In subsection 4.1 the multilevel setting will be introduced, in which each level has its own cLIS, and these converge as the level increases. The form of the importance sampling proposal will be described in section 4.2. The embedding

of cLIS will be described in subsection 4.3, and the ML covariance construction using the cLIS will be described in section 4.4. The additional ML cost considerations due to the DILI mutations are considered in subsection 4.5, and finally an example of the framework for a simple basis is presented in subsection 4.6.

4.1. Setting. Recall that in the setting of section 2, we are interested in a sequence of unnormalized densities $\kappa_\ell(u_{0:\ell})$ in (2.1) defined on spaces of increasing dimension E_ℓ for levels $\ell = 0, 1, \ldots$. Let h_ℓ denote a resolution parameter and $\mathcal{C}_\ell$ the associated computational cost of evaluating $\kappa_\ell(u_{0:\ell})$, such that $h_\ell \to 0$ and $\mathcal{C}_\ell \to \infty$ as $\ell \to \infty$, and assume that the computational cost is dominated by a forward model involved in the likelihood calculation $\mathcal{L}_\ell(u_{0:\ell})$, as in (2.1). In particular, consider the case in which the sequence of spaces $E_0, E_1, \ldots, E_L$ correspond to finite-dimensional approximations (of increasing dimension) of a limiting space $E := E_\infty$, where E is a separable Hilbert space, and $u \in E$.

In order to establish a clear context, let $\phi_1, \phi_2, \ldots \in E$ and define $\Psi_\ell := [\phi_1, \ldots, \phi_{d_\ell}] \in E \times \mathbb{R}^{d_\ell}$. Using matrix notation, let

$$E_\ell = (\Psi_\ell^\top \Psi_\ell)^{-1} \Psi_\ell^\top E.$$

Letting $u_{0:\ell} = (\Psi_\ell^\top \Psi_\ell)^{-1} \Psi_\ell^\top u$, $\Psi_\ell u_{0:\ell}$ is the orthogonal projection of u onto the d_ℓ-dimensional subspace of E spanned by the columns of Ψ_ℓ. In the following $u_{0:\ell}$ may also correspond to the value of u at d_ℓ grid points with $\Psi_\ell u_{0:\ell}$ an interpolant through those points. In the limit, isomorphic representations of E will be identified, i.e., spatial representations or sequence representations in terms of expansion coefficients. Suppose that one has a regularly structured grid that is uniform across D underlying spatiotemporal dimensions of the limiting space E (for example, $L^2([0,1]^D, \mathbb{R})$) and that the grid spacing is h_ℓ. Then the dimension of E_ℓ is $d_\ell = h_\ell^{-D}$. Conversely, for an arbitrary expansion, for instance, in terms of some family of orthonormal polynomials, with equal numbers of basis functions in each direction, it is reasonable to *define* $h_\ell := d_\ell^{-1/D}$. These notions are therefore interchangeable.

Let $P_\ell \in \mathbb{R}^{d_\ell \times m_\ell}$ denote an orthonormal basis for the m_ℓ-dimensional cLIS at level ℓ, so that

$$C_\ell = I_{d_\ell} - Q_\ell Q_\ell^\top,$$

where $Q_\ell = P_\ell \Lambda_\ell^{1/2}$ for some diagonal matrix Λ_ℓ of nonzero singular values, and I_{d_ℓ} is the $d_\ell \times d_\ell$ identity matrix. We set $m = \lim_{\ell \to \infty} m_\ell$ and let P denote the limiting m-dimensional cLIS on E.

According to the simulated examples, the cLIS associated with E_ℓ is expected to require $\mathcal{O}(d_\ell)$ samples to identify—see Figures 1 and 2(b). Therefore we cannot afford to compute the cLIS at level L, or high levels close to L, without affecting the cost of the algorithm. It is reasonable to assume that for ℓ sufficiently large, $m_\ell \approx m$, and hence we can obtain a good approximation of P. Therefore, at some level ℓ^* in the MLSMC algorithm, one stops constructing the cLIS and the current cLIS $P_{\ell^*} \subset E_{\ell^*}$ is simply embedded into E_{ℓ^*+n} for $n \geq 1$. Thus, one can use the empirical covariance on the cLIS, at a cost that depends upon m (at most cubically), for a DILI proposal without recomputing the cLIS on higher levels. The proposal construction therefore does not depend upon L or ε except through the forward model, and hence it does not impact the asymptotic cost. Furthermore, within the MLSMC

context, one needs to collect at least d_ℓ samples for $\ell < \ell^*$, but the restriction does not persist for $\ell > \ell^*$. The implication of this is discussed in more detail in section 4.5.

4.2. Importance sampling proposal to extend dimension. The mutation kernel K_ℓ will be constructed through the DILI methodology in section 3.3. It remains to determine the kernel $q_\ell : E_{\ell-1} \to \mathcal{P}(U_\ell)$ that extends the dimension of the state space during the iterative importance sampling steps. In both numerical applications in section 5 we employ regular grids of increasing resolution in one dimension and two dimensions; other options could involve truncating the Karhunen–Loève expansion of the prior Gaussian measure.

In our applications we have used the Gaussian prior dynamics to determine q_ℓ, so that

$$q_\ell(u_{0:\ell-1}, du_\ell) = \mu_0(du_\ell | u_{0:\ell-1}),$$

though other choices could also be made. This choice gives

$$G_\ell(u_{0:\ell}) = \mathcal{L}_\ell(u_{0:\ell}) / \mathcal{L}_{\ell-1}(u_{0:\ell-1}).$$

From standard properties of Gaussian laws, assuming that $\mu_0(du_{0:\ell}) = N(0, \Gamma_{0:\ell})$ with covariance

$$\Gamma_{0:\ell} = \begin{pmatrix} \Gamma_{0:(\ell-1)} & \Gamma_{0:(\ell-1),\ell} \\ \Gamma_{0:(\ell-1),\ell}^\top & \Gamma_{\ell,\ell} \end{pmatrix}$$

with $\Gamma_{0:(\ell-1)} \in \mathbb{R}^{d_{\ell-1} \times d_{\ell-1}}$, $\Gamma_{\ell,\ell} \in \mathbb{R}^{d'_\ell \times d'_\ell}$, $\Gamma_{0:(\ell-1),\ell} \in \mathbb{R}^{d_{\ell-1} \times d'_\ell}$, we have

$$q_\ell(u_{0:\ell-1}, \cdot) = \Gamma_{0:(\ell-1),\ell}^\top \Gamma_{0:(\ell-1)}^{-1} u_{0:\ell-1} + N\big(0, \Gamma_\ell\big), \tag{4.1}$$

where

$$\Gamma_\ell := \Gamma_{\ell,\ell} - \Gamma_{0:(\ell-1),\ell}^\top \Gamma_{0:(\ell-1)}^{-1} \Gamma_{0:(\ell-1),\ell}\,.$$

4.3. cLIS construction when extending dimension. Recall that the main idea in section 4.1 is that one will reach a cut-off level, say, $\ell - 1$, when the standard cLIS methodology will be applied using the particle information available at this point, as described in section 3.3, but from level ℓ onward the cLIS will simply be propagated forward without Monte Carlo effort to identify further directions informed by the likelihood. We will now describe how to carry out this propagation.

The construction of the cLIS proposal in section 3.3 requires the identification of an orthonormal set of vectors spanning the critical subspace informed by the likelihood after whitening the prior covariance. That is, one must in practice work with the linear transformation $v_{0:\ell} = L_\ell^{-1} u_{0:\ell}$, where L_ℓ is any matrix such that $L_\ell L_\ell^\top = \Gamma_{0:\ell}$. Notice that in many cases (e.g., if the prior is a Gaussian Markov random field) L_ℓ^{-1} is sparse, so this operation is cheap. Also, L_ℓ itself may be sparse, or have a simple structure which allows for cheap (i.e., not $\mathcal{O}(d_\ell^2)$) operations, as will be the case in section 5.1 below. Assume that the columns of matrix $P_{\ell-1} \in \mathbb{R}^{d_{\ell-1} \times m}$ correspond to the orthonormal basis of the cLIS at the cut-off level $\ell - 1$, so that $P_{\ell-1}^\top P_{\ell-1} = I_m$ and

$$L_{\ell-1}^{-1} C_{\ell-1} L_{\ell-1}^{-\top} = I_{d_{\ell-1}} - P_{\ell-1} \Lambda P_{\ell-1}^\top.$$

We will identify P_ℓ.

It will be convenient to define the matrices

$$A_{\ell|\ell-1} = \begin{pmatrix} I_{d_{\ell-1}} \\ \Gamma_{0:(\ell-1),\ell}^{\top}\Gamma_{0:(\ell-1)}^{-1} \end{pmatrix}, \qquad A_{\ell\backslash\ell-1} = \begin{pmatrix} 0_{d_{\ell-1}\times d'_\ell} \\ \Gamma_\ell^{1/2} \end{pmatrix}, \tag{4.2}$$

where $I_{d_{\ell-1}}$ is the $d_{\ell-1}\times d_{\ell-1}$ identity matrix and $0_{d_{\ell-1}\times d'_\ell}$ is the $d_{\ell-1}\times d'_\ell$ matrix of all zeros. Then for $u_\ell^* \sim q_\ell(u_{0:\ell-1},\cdot)$, one has

$$(u_{0:\ell-1}^{\top}, u_\ell^{*,\top})^{\top} = A_{\ell|\ell-1}u_{0:\ell-1} + A_{\ell\backslash\ell-1}\xi_\ell\,,$$

where $\xi_\ell \sim N(0, I_{d'_\ell})$. Indeed, by definition we have

$$\Gamma_{0:\ell} = A_{\ell|\ell-1}\Gamma_{0:\ell-1}A_{\ell|\ell-1}^{\top} + A_{\ell\backslash\ell-1}A_{\ell\backslash\ell-1}^{\top}, \tag{4.3}$$

so that if $u_{0:\ell-1}\sim\mu_0$, then $(u_{0:\ell-1}^{\top}, u_\ell^{*,\top})^{\top}\sim\mu_0$.

In other words, $A_{\ell|\ell-1}$ has rank $d_{\ell-1}$, and its column-space is exactly the $d_{\ell-1}$-dimensional subspace of $\mathbb{R}^{d_\ell}$ in which $u_{0:\ell}$ depends upon $u_{0:\ell-1}$, under the prior measure μ_0. Therefore, if one has a genuine (rather than approximate) cLIS and it is furthermore already completely characterized at level $\ell-1$, then $A_{\ell|\ell-1}$ embeds it into level ℓ, but with respect to the $u_{0:\ell}$ variables. The appropriate cLIS with respect to $u_{0:\ell-1}$ is $L_{\ell-1}P_{\ell-1}$, orthogonal with respect to the Mahalanobis norm weighted with $L_{\ell-1}^{-\top}L_{\ell-1}^{-1} = \Gamma_{0:\ell}^{-1}$. Therefore one expects $L_\ell P_\ell = A_{\ell|\ell-1}L_{\ell-1}P_{\ell-1}$ to be the appropriate cLIS on $u_{0:\ell}$ embedded into level ℓ. In other words,

$$P_\ell = L_\ell^{-1}A_{\ell|\ell-1}L_{\ell-1}P_{\ell-1}\,. \tag{4.4}$$

A useful identity is the following:

$$\Gamma_{0:\ell-1}^{-1} = A_{\ell|\ell-1}^{\top}\Gamma_{0:\ell}^{-1}A_{\ell|\ell-1}\,. \tag{4.5}$$

To see this, observe that the first $d_{\ell-1}$ rows of $\Gamma_{0:\ell}$ are given by $\Gamma_{0:\ell-1}A_{\ell|\ell-1}^{\top}$. It is then clear that

$$A_{\ell|\ell-1}^{\top}\Gamma_{0:\ell}^{-1} = \Gamma_{0:(\ell-1)}^{-1}\begin{pmatrix} I_{d_{\ell-1}} & 0_{d_{\ell-1}\times d'_\ell}\end{pmatrix},$$

and the identity above follows immediately. Due to (4.5) it is easy to check that P_ℓ defined by (4.4) satisfies $P_\ell^{\top}P_\ell = I_m$. This ensures that an orthonormal cLIS at the cut-off level $\ell-1$ transforms to an orthonormal cLIS at level ℓ through the map $P_{\ell-1}\mapsto L_\ell^{-1}A_{\ell|\ell-1}L_{\ell-1}P_{\ell-1}$.

4.4. Multilevel covariance estimation. The covariance C_ℓ can also be estimated with the multilevel estimator [5, 22]

$$C_\ell^{\mathrm{ML}} \approx C_0^{N_0} + \sum_{l=1}^{\ell}\left(C_l^{N_l} - C_{l-1}^{N_l}\right), \tag{4.6}$$

where $C_l^{N_l} = \frac{1}{N_l}\sum_{i=1}^{N_l} u_{0:l}^i(l)(u_{0:l}^i(l))^{\top} - \left(\frac{1}{N_l}\sum_{i=1}^{N_l}u_{0:l}^i(l)\right)\left(\frac{1}{N_l}\sum_{i=1}^{N_l}u_{0:l}^i(l)\right)^{\top}$, and $C_{l-1}^{N_l}$ is the appropriate upscaled (so that matrix dimensions match in 4.6)) sample covariance associated

with $u_{0:l-1}(l)$. This will give rise to the multilevel cLIS approximation P_ℓ^{ML}, which will be used to approximate the covariance on the approximate cLIS, $\Sigma_\ell = P_\ell^{\mathrm{ML},\top} C_\ell P_\ell^{\mathrm{ML}}$, by

$$\Sigma_\ell^{\mathrm{ML}} \approx P_\ell^{\mathrm{ML},\top} C_0^{N_0} P_\ell^{\mathrm{ML}} + \sum_{l=1}^{\ell} P_\ell^{\mathrm{ML},\top} \left(C_l^{N_l} - C_{l-1}^{N_l} \right) P_\ell^{\mathrm{ML}}.$$

Consider $A_{\ell+n|\ell} = A_{\ell+n|\ell+n-1} A_{\ell+n-1|\ell+n-2} \cdots A_{\ell+1|\ell}$ for $A_{l|l-1}$, $l \geq 1$, defined in (4.2). As mentioned in section 4.3, the cLIS will be constructed only until some level ℓ^*. Then the cLIS $P_{\ell^*}^{\mathrm{ML}} \in \mathbb{R}^{d_{\ell^*} \times m}$, constructed at the final level using (4.6), is transformed into the cLIS at higher levels $P_{\ell^*+n}^{\mathrm{ML}} \in \mathbb{R}^{d_{\ell^*+n} \times m}$ by

$$P_{\ell^*+n}^{\mathrm{ML}} = L_{\ell^*+n}^{-1} A_{\ell^*+n|\ell^*} L_{\ell^*} P_{\ell^*}^{\mathrm{ML}}, \tag{4.7}$$

where orthonormality of the column vectors of $P_{\ell^*+n}^{\mathrm{ML}}$ holds by transitivity and (4.5).

4.5. Multilevel cost considerations. In the following discussion the constants β and γ are defined in Appendix A, assumption (A3). The multilevel analysis proceeds as in a standard case, except one has to consider that if $\ell > \ell^*$, then $\mathcal{C}_\ell = h_\ell^{-\gamma D}$, and if $\ell \leq \ell^*$, then $\mathcal{C}_\ell = h_\ell^{-3D}$. Note that the cubic power corresponds to the worst case scenario for the cost of computing the cLIS, while it may be possible in cases to compute it more cheaply, e.g., even with linear cost. Assuming we fix ℓ^*, then asymptotically the use of the cLIS does not change the error analysis of the estimates. One has $N_\ell = h_\ell^{(\beta+\gamma)/2}$ if $\ell > \ell^*$ and $N_\ell = h_\ell^{(\beta+3)/2}$ if $\ell \leq \ell^*$. More careful analysis can be done, e.g., using the rate of convergence of the cLIS to choose ℓ^*, but since this construction is merely to improve mixing of the MCMC kernels, it is reasonable to simply fix ℓ^* and ensure that N_0 is chosen large enough so that $N_{\ell^*} > d_{\ell^*}$.

4.6. Example with a Karhunen–Loève basis. We end this subsection with a comment that the spaces $\{E_\ell\}_{\ell=0}^L$ could be determined via a Karhunen–Loève expansion as described below.

Let μ_0 be the prior distribution over the infinite-dimensional separable Hilbert space E, which will be assumed Gaussian with mean 0 and trace-class covariance operator Γ. There is an orthonormal basis $\{\phi_i\}_{i=1}^\infty$ for E and associated eigenvalues $\{\lambda_i\}_{i=1}^\infty$ such that $\Gamma\phi_i = \lambda_i \phi_i$. The Karhunen–Loève expansion of a draw $u \sim \mu_0$ is given by

$$u = \sum_{i=1}^{\infty} x_i \phi_i, \quad \text{where} \quad x_i = \langle u, \phi_i \rangle = \lambda_i^{1/2} \xi_i, \quad \text{and} \quad \xi_i \sim N(0,1) \ \text{ i.i.d.} \tag{4.8}$$

Thus, the covariance operator Γ is diagonal in the basis $\Psi_\infty = [\phi_1, \phi_2, \ldots]$. In this setting it is natural to work with the coordinates $u_\ell = (x_{d_{\ell-1}+1}, \ldots, x_{d_\ell})$, $0 \leq \ell \leq L$, so that we simply have $L_\ell = \Gamma_{0:\ell}^{1/2} = \Psi_\ell^\top \Gamma^{1/2} \Psi_\ell = \mathrm{diag}\{\lambda_1^{1/2}, \ldots, \lambda_{d_\ell}^{1/2}\}$. Also, one has that

$$q_\ell(u_{0:\ell-1}, \cdot) = q_\ell(\cdot) = N\left(0, \mathrm{diag}(\lambda_{d_{\ell-1}+1}, \lambda_{d_{\ell-1}+2}, \ldots, \lambda_{d_\ell})\right), \tag{4.9}$$

and $\Gamma_{0:(\ell-1),\ell} = 0$ so $P_{\ell^*+\ell} = [P_{\ell^*}^\top, 0_{m_{\ell^*} \times (d_\ell - d_{\ell^*})}]^\top$, where recall that $0_{m\times n} \in \mathbb{R}^{m\times n}$ is a matrix of zeros.

5. Examples. In this section, two models will be described. Section 5.1 considers inversion of the white noise forcing in an SDE given noisy observations of the path. Section 5.2 considers a Bayesian inverse problem of inferring the diffusion coefficient in a 2D elliptic PDE given noisy observations of the solution field. Some theoretical considerations relating to verification of the assumptions (in Appendix A) for the algorithms are included in Appendix B.

5.1. Conditioned diffusions. We consider an SDE scenario. For u denoting a realization of the s-dimensional Brownian motion, $s \geq 1$, let $p = p(u)$ be the solution of the SDE

$$dp = f(p)dt + \sigma(p)du, \quad p(0) = p_0, \tag{5.1}$$

where $f : \mathbb{R}^s \mapsto \mathbb{R}^s$, $\sigma : \mathbb{R}^s \mapsto \mathbb{R}^{s\times s}$ are elementwise Lipschitz continuous with $\sigma \in \mathbb{R}^{s\times s}$ nondegenerate. Let $\mathcal{G}_i(u) = p(t_i; u)$ for times $0 < t_1 < \cdots < t_q \leq T$, $q \geq 1$. We consider observations $y|u \sim N(\mathcal{G}(u), \Xi)$, where $\mathcal{G} = (\mathcal{G}_1^\top, \ldots, \mathcal{G}_q^\top)^\top$, with noise (of variance Ξ) independent of u, so that the likelihood is

$$\mathcal{L}(u) = \exp\left(-\tfrac{1}{2}|y - \mathcal{G}(u)|_\Xi^2\right). \tag{5.2}$$

5.1.1. Numerical method and multilevel approximation. We will henceforth assume $s = 1$, though multidimensional extensions are straightforward. The standard Euler–Maruyama discretization is employed, with refinement occurring via Brownian bridge sampling between successive grid points; this is a particular scenario of the general description in section 4.2. The paths are generated on a uniform grid, which gives rise to proposals of the form (4.1) under the prior Wiener measure dynamics. In particular, let us assume that $d_\ell = d_0\, 2^\ell$, so $h_\ell = T/d_\ell$; to avoid undue complications d_0 is chosen large enough to accommodate the q observations at grid points. Then the linear transformations in (4.1) are given for the case of the scalar ($s = 1$) SDE in (5.1) by the following, for $i = 1, \ldots, d_\ell$ (the first, undefined, equation is ignored for $i = 1$):

$$\begin{aligned}(A_{\ell+1|\ell})_{2i-1,i-1} &= (A_{\ell+1|\ell})_{2i-1,i} = 1/2,\\ (A_{\ell+1|\ell})_{2i,2i} &= 1,\\ (A_{\ell+1\backslash\ell})_{2i-1,i} &= \sqrt{h_\ell}/2,\end{aligned}$$

and $(A_{\ell+1|\ell})_{j,k} = (A_{\ell+1\backslash\ell})_{j,k} = 0$ otherwise. This is simply a way to write down the well-known Brownian bridge measure for the fine grid points $u_{0:\ell+1}$, every other of which coincides with one of the coarse grid points $u_{0:\ell}$ or bisects two of them. The new bisecting points $u_{\ell+1}$ are conditionally independent given $u_{0:\ell}$, with distribution,

$$q_{\ell+1,i}(u_{0:\ell}, u_{\ell+1,i}) = N\left(\tfrac{1}{2}(u_{0:\ell,i} + u_{0:\ell,i+1}), \tfrac{h_\ell}{4}\right)$$

for $i = 1, \ldots, d'_\ell$. Operator $L_\ell : v_{0:\ell} \mapsto u_{0:\ell}$ in (4.5) is given by the Cholesky factorization: $(L_\ell)_{j,i} = \sqrt{h_\ell}$, $i \leq j$; $(L_\ell)_{j,i} = 0$ otherwise.

5.1.2. Numerical results. The specific settings of our numerical study are as follows: $\sigma(p) = 1$, $T = q = 16$, and the observations are evenly spaced with $t_1 = 1$ and noise $\Xi = 0.01$. The simulations are carried out with $d_0 = 32$ at the initial level and $d_l = d_0 2^l$ as described above. For simplicity the quantity of interest is taken as the observation function $\varphi = \mathcal{G}$.

Numerical results for the solution of the conditioned diffusion problem are shown in Figure 3. The variance rate plot helps us to obtain β for our simulations. In Appendix A, assumption (A3), a precise definition of β is given, but this is essentially associated to the variance, where larger β relates to smaller variances. We then consider SMC (i.e., no telescoping identity), MLSMC with the standard pCN method for the mutations and MLSMC with the DILI proposals of section 3. The samples for the simulations are chosen as mentioned above. The results are repeated 100 times and averaged for robustness. The (theoretical) cost vs. error (mean square error) plot of Figure 3 presents a comparison of the three methods. Both MLSMC methods outperform SMC as was the case in [4]. Moreover, it is evident that the performance with the DILI mutations is superior to that of the standard pCN mutations.

5.2. Elliptic PDE inverse problem. In this section, we consider a Bayesian inverse problem involving inference of the (log) permeability coefficient in a 2D elliptic PDE, given noisy measurements of the associated solution field (representing, e.g., pressure). Consider the nested spaces $V := H^1(\Omega) \subset L^2(\Omega) \subset H^{-1}(\Omega) =: V^*$ for a domain $\Omega \subset \mathbb{R}^2$ with convex boundary $\partial\Omega \in C^0$. Let $f \in V^*$, and consider the following PDE on Ω:

$$-\nabla \cdot (\mathbb{K}(u)\nabla p) = f \qquad \text{on } \Omega, \tag{5.3}$$

$$p = 0 \qquad \text{on } \partial\Omega \tag{5.4}$$

for pressure field p, permeability $\mathbb{K}(u) = e^u$, and known force field f. We set up a Bayesian inference problem for the unknown log permeability field u. We assume a truncated stationary Gaussian prior,

$$u \sim \mu_0(du) \cdot \mathrm{I}[\,|u|_\infty < R\,], \quad \mu_0 \equiv N(0, C), \tag{5.5}$$

for some $R > 0$ with C denoting the covariance operator derived through the covariance function

$$c(x, x') = \sigma^2 \exp\{-|x - x'|^2/\alpha\} \tag{5.6}$$

for hyperparameters $\sigma > 0$, $\alpha > 0$.

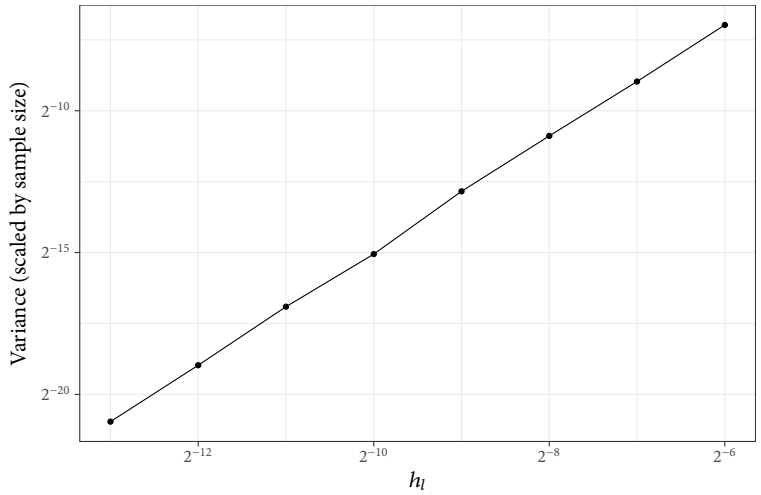
We will henceforth assume $\Omega = [0, 1]^2$. Let $p(\,\cdot\,; u)$ denote the weak solution of (5.3)–(5.4) for parameter u. Define the following vector-valued function:

$$\mathcal{G}(u) = [g_1(p(\,\cdot\,; u)), \ldots, g_M(p(\,\cdot\,; u))]^\top,$$

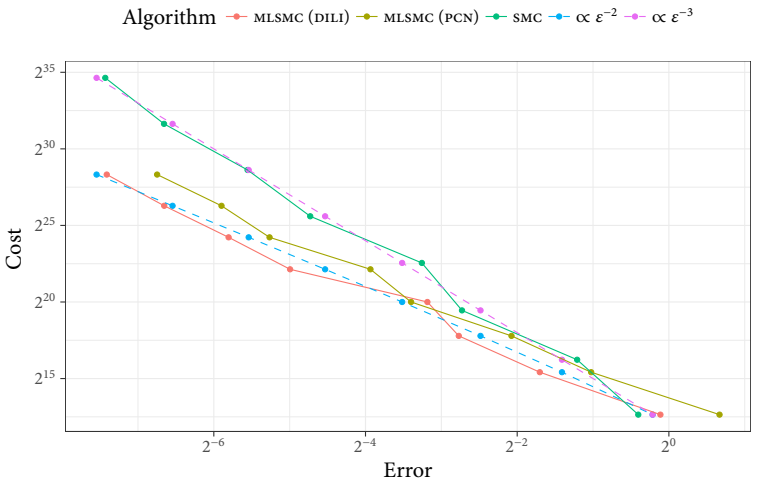
where g_m are elements of the dual space V^* for $m = 1, \ldots, M$ for some $M > 1$. It is assumed that the data take the form

$$y = \mathcal{G}(u) + \xi, \qquad \xi \sim N(0, \Xi), \qquad \xi \perp u, \tag{5.7}$$

so that the likelihood is given again by $\mathcal{L}(u) = \exp(-\frac{1}{2}|y - \mathcal{G}(u)|_\Xi^2)$.



(a) Variance convergence rate.



(b) Cost vs. error.

Figure 3. *Results for the conditioned diffusion example.*

5.2.1. Numerical method and multilevel approximation. Consider the 1D piecewise linear nodal basis functions ϕ_j^1 defined as follows for mesh $\{x_i = i/K\}_{i=0}^K$:

$$\phi_j^1(x) = \begin{cases} \frac{x-x_{j-1}}{x_j-x_{j-1}}, & x \in [x_{j-1}, x_j], \\ 1 - \frac{x-x_j}{x_{j+1}-x_j}, & x \in [x_j, x_{j+1}], \\ 0 & \text{otherwise}. \end{cases}$$

Consider the tensor product grid over $\Omega = [0,1]^2$ formed by $\{(x_i, x_j)\}_{i,j=1}^K$, where $K = K_0 \times 2^\ell$ with initial resolution $K_0 = 10$. Let $\phi_{i,j}(x,y) = \phi_j^1(x)\phi_i^1(y)$ be piecewise bilinear functions, and

let $E_\ell = \text{span}\{\phi_i; 1 \le i \le d_\ell\}$ with $d_\ell = K^2$, and any appropriate single index representation. The permeability at level ℓ will be approximated by $\mathbb{K}_\ell(u_{0:\ell}) = \sum_{i=1}^{d_\ell} e^{u^i_{0:\ell}}\phi_i$. Likewise, the solution will be approximated by $p_\ell(u_{0:\ell}) = \sum_{i=1}^{d_\ell} p^i_\ell\,\phi_i$. The weak solution of the considered PDE (5.3)–(5.4) is generated by a standard finite element approximation, resulting in the solution $\mathbf{p} := p_\ell(u_{0:\ell})$. This is done by substituting these expansions into (5.3) and taking inner product with ϕ_j for $j = 1, \dots, d_\ell$. Define $\mathbf{f}_j = \langle f, \phi_j\rangle$ and

$$\mathbf{A}_{ij} := \sum_{k_1=j_1-1}^{j_1+1}\sum_{k_2=j_2-1}^{j_2+1}\int_{x_{j_1-1}}^{x_{j_1+1}}\int_{y_{j_2-1}}^{y_{j_2+1}} e^{u^k_{0:\ell}}\phi_k\nabla\phi_i\cdot\nabla\phi_j dxdy,$$

where the notation $j = (j_1, j_2)$ is introduced to represent the components of the indices corresponding to spatial dimensions 1 and 2. The approximate weak solution to (5.3), (5.4) is given by the system $\mathbf{Ap} = \mathbf{f}$.

The solution $p_\ell(u_{0:\ell})$ is then plugged into the likelihood to provide $\mathcal{L}_\ell(u_{0:\ell})$. At the next level, values of log-permeability on extra grid points are proposed from the conditional prior dynamics $u_{\ell+1}|u_{0:\ell}$, by halving horizontal/vertical distances between points in the grid.

5.2.2. Numerical results. The specific settings for our simulations and generated data are as follows: the source/sink term f is defined by a superposition of four weighted Gaussian bumps with standard deviation $\sigma_f = 0.05$, centered at $(0.3, 0.3)$, $(0.3, 0.7)$, $(0.7, 0.3)$, and $(0.7, 0.7)$, with weights $\{2, -3, -2, 3\}$, respectively. Observations of the potential function p are collected at 25 measurement points, evenly spaced within $[0.2, 0.6]^2$ (boundaries included). The observation variance σ_y^2 is chosen such that a prescribed signal-to-noise ratio, which is defined as $\max\{p\}/\sigma_y$, is equal to 10. The hyperparameters α and σ^{-2} are given Gamma priors with mean and variance 1. For simplicity the quantity of interest is taken as the observation function $\varphi = \mathcal{G}$.

The numerical results for the elliptic PDE inverse problem are presented in Figure 4, which contains plots analogous to those shown for the previous numerical example. Again, the MLSMC schemes show the desired improved convergence rate, and the DILI mutation steps yield consistently better performance than pCN mutations.

Appendix A. Basic theoretical results.

The following assumptions will be made. Throughout E_ℓ is compact for any fixed $\ell < +\infty$.

(A1) Assume there exist some $\underline{c}$, $\overline{C}$ such that for all $\ell = 0, 1, \dots,$ and all $u_{0:\ell} \in E_\ell$

$$0 < \underline{c} \le G_\ell(u_{0:\ell}) \le \overline{C} < \infty. \tag{A.1}$$

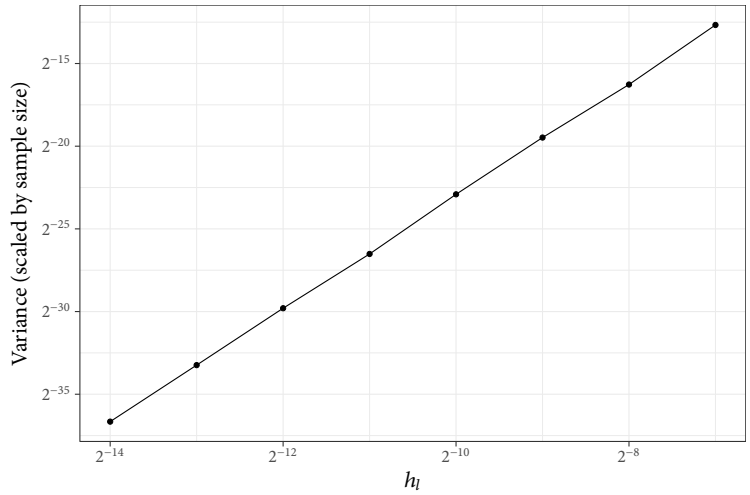
(A2) Assume there exists a $\lambda < 1$ such that for all $\ell = 0, 1, \dots,$ $u, v \in E_\ell$ and $A \subset E_\ell$

$$K_\ell(u, A) \ge \lambda K_\ell(v, A).$$

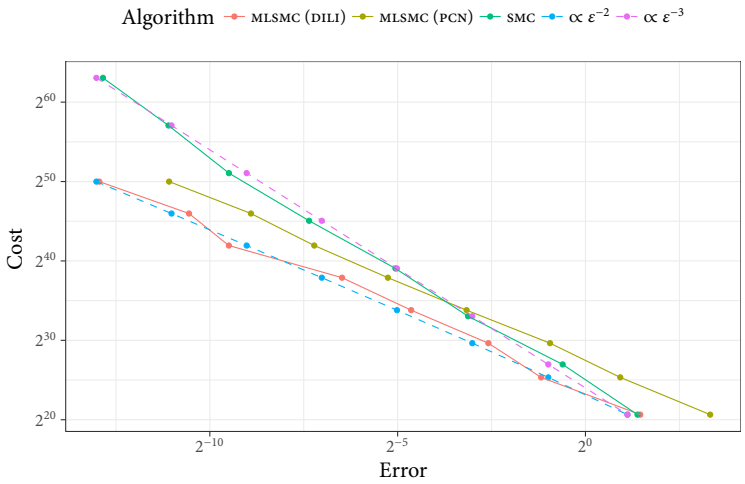
(A3) Assume there exists a $c > 0$ and $\beta > 0$ such that for all ℓ sufficiently large

$$V_\ell := \max\{\|G_\ell - 1\|^2_\infty, \|\rho_\ell - \rho_{\ell-1}\|^2_\infty\} \le c h_\ell^\beta, \tag{A.2}$$

where for bounded and measurable $f : E_\ell \to \mathbb{R}^d$, $\|f\|_\infty = \sup_{u\in E_\ell}|f(u)|$, $|\cdot|$ is the L_1-norm, and h_ℓ denotes an accuracy parameter, for example, mesh-diameter of the



(a) Variance convergence rate.



(b) Cost vs. error.

Figure 4. *Results for the 2D PDE example.*

discretization of a PDE. Also, assume the cost $\mathcal{C}_\ell$ to evaluate G_ℓ and ρ_ℓ satisfies, for some $\zeta \geq 0$,

$$\mathcal{C}_\ell \leq c\, h_\ell^{-\zeta}.$$

Define

$$\hat{\eta}_L^{\mathrm{ML}}(\varphi) := \eta_0^{N_0}(\varphi \circ \rho_0) + \sum_{\ell=0}^{L} \eta_\ell^{N_\ell}(G_\ell)^{-1} \eta_\ell^{N_\ell}(G_\ell\, \varphi \circ \rho_\ell) - \eta_\ell^{N_\ell}(\varphi \circ \rho_{\ell-1}).$$

Let $a(\epsilon) \lesssim b(\epsilon)$ denote that there exists a $c > 0$ such that $a(\epsilon) \lesssim cb(\epsilon)$ for all ϵ sufficiently small.

Proposition A.1. *Assume* (A1)–(A3) *and that* $\beta > \zeta$. *Then, for any* $\varepsilon > 0$ *there exists an* L, *and a choice of* $\{N_\ell\}_{\ell=0}^L$, *such that*

$$\mathbb{E}|\hat{\eta}_L^{\mathrm{ML}}(\varphi) - \hat{\eta}(\varphi)|^2 \lesssim \varepsilon^2 \tag{A.3}$$

for a total cost $\mathrm{Cost} \lesssim \varepsilon^{-2}$.

Proof. The proof follows essentially that of [4] given the above assumptions. Observe that Lemma A.3 (in Appendix A.1) below provides the bound

$$\mathbb{E}\left[\{\hat{\eta}_L^{\mathrm{ML}}(g) - \mathbb{E}_{\eta_L}[g(U)]\}^2\right] \leq C\left(\frac{1}{N_0} + \sum_{\ell=1}^{L} \frac{V_\ell}{N_\ell} + \sum_{1\leq \ell < q \leq L} V_\ell^{1/2} V_q^{1/2} \left\{\frac{\kappa^q}{N_\ell} + \frac{1}{N_\ell^{1/2} N_q}\right\}\right).$$

Theorem 3.3 of [24] describes how to complete the proof. Briefly, the choice $L \eqsim |\log \varepsilon|$ controls the bias. One chooses $N_\ell = \varepsilon^{-2} K_L h_\ell^{(\beta+\zeta)/2}$, where $K_L = \sum_{\ell=1}^{L-1} h_\ell^{(\beta-\zeta)/2} = \mathcal{O}(1)$, so one has

$$\mathrm{COST} = \sum_{\ell=0}^{L} N_\ell C_\ell = \varepsilon^{-2} K_L^2 \lesssim \varepsilon^{-2}.$$

It then suffices to show the second term is negligible for this choice, and this is done in Theorem 3.3 of [24]. ∎

The below result follows directly from that in [14] and hence the proof is omitted.

Corollary A.2. *Assume* (A1)–(A3) *and assume* $\beta > \zeta$. *Then, for any* $\varepsilon > 0$ *there exists an* L *and a choice of* $\{N_\ell\}_{\ell=0}^L$ *such that*

$$\mathbb{E}|\gamma_L^{N_{0:L}}(G_L) - Z_\infty / Z_0|^2 \lesssim \varepsilon^2 \tag{A.4}$$

for a total cost $\mathrm{Cost} \lesssim |\log \varepsilon| \varepsilon^{-2}$.

A.1. Key theoretical result. The following lemma is similar to Theorem 3.1 in [4], and the proof follows in the same spirit, but is given for completeness.

Lemma A.3. *Assume* (A1)–(A3). *Then there exists a* $C > 0$ *and* $\kappa \in (0, 1)$ *such that for any* $g \in \mathcal{B}_b(E)$ *with* $\|g\|_\infty = 1$,

$$\mathbb{E}\left[\{\hat{\eta}_L^{\mathrm{ML}}(g) - \mathbb{E}_{\eta_L}[g(U)]\}^2\right] \leq C\left(\frac{1}{N_0} + \sum_{l=1}^{L} \frac{V_l}{N_l} + \sum_{1\leq l < q \leq L} V_l^{1/2} V_q^{1/2} \left\{\frac{\kappa^q}{N_l} + \frac{1}{N_l^{1/2} N_q}\right\}\right).$$

Proof. The proof follows essentially that of [4] given the above assumptions. Assumptions (A1)–(A2) are similar to that paper. Note that, as shown in section 4.2 of [4], there is a constant $C > 0$ such that

$$\|\tfrac{Z_{l-1}}{Z_l} G_l - 1\|_\infty \leq C \|G_l - 1\|_\infty. \tag{A.5}$$

Observe that $\hat{\eta}_l(\varphi \circ \rho_l) = \eta_l(G_l\varphi \circ \rho_l)/\eta_l(G_l)$. Now establish the following notation:

$$Y_l^{N_l} = \frac{\eta_l^{N_l}(G_l\varphi \circ \rho_l)}{\eta_l^{N_l}(G_l)} - \eta_l^{N_l}(\varphi \circ \rho_{l-1}) \ ,$$

$$Y_l = \frac{\eta_l(G_l\varphi \circ \rho_l)}{\eta_l(G_l)} - \eta_l(\varphi \circ \rho_{l-1}) \quad \big(\equiv \eta_l(g) - \eta_{l-1}(g) \big) \ , \tag{A.6}$$

$$\overline{\varphi_l}(u) = \left(\frac{Z_{l-1}}{Z_l} G_l(u) - 1 \right) \ ,$$

$$\widetilde{\varphi}_l(u) = \overline{\varphi_l}(u)\varphi(u) \ ,$$

$$A_n(\varphi, N) = \eta_n^N(G_n\varphi \circ \rho_n)/\eta_n^N(G_n) \ , \quad \varphi \in \mathcal{B}_b(E) \ , \quad 0 \le n \le L-1 \ , \tag{A.7}$$

$$\overline{A}_n(\varphi, N) = A_n(\varphi, N) - \frac{\eta_n(G_n\varphi \circ \rho_n)}{\eta_n(G_n)} \ . \tag{A.8}$$

Notice that $\eta_l(\overline{\varphi_l}) = 0$ and $\eta_l(G_l) = Z_l/Z_{l-1}$. So,

$$Y_l^{N_l} - Y_l = \underbrace{A_l(\varphi, N_l)\,\{\eta_l - \eta_l^{N_l}\}(\overline{\varphi_l})}_{T_l^1} + \underbrace{\{\eta_l^{N_l} - \eta_l\}(\widetilde{\varphi}_l \circ \rho_l)}_{T_l^2} + \underbrace{\{\eta_l^{N_l} - \eta_l\}(\varphi \circ (\rho_l - \rho_{l-1}))}_{T_l^3} \ . \tag{A.9}$$

Observe that there is an additional term T_l^3 in comparison to equation (10) of [4]. Lemma 3.1 of that paper is replaced by

$$\begin{aligned} \|Y_l^{N_l} - Y_l\|_2^2 &\le 4\|A_l(\varphi, N_l)\,\{\eta_l - \eta_l^{N_l}\}(\overline{\varphi_l})\|_2^2 \\ &\quad + 4\|\{\eta_l^{N_l} - \eta_l\}(\widetilde{\varphi}_l \circ \rho_l)\|_2^2 + 4\|\{\eta_l^{N_l} - \eta_l\}(\varphi \circ (\rho_l - \rho_{l-1}))\|_2^2. \end{aligned} \tag{A.10}$$

In view of (A.5) and [11, Theorem 7.4.4], the first two terms are bounded by $C\|G_l - 1\|_\infty^2/N_l$ and the last term is bounded by $C\|\rho_l - \rho_{l-1}\|_\infty^2/N_l$. Now

$$\mathbb{E}\Big[\Big\{\sum_{l=1}^N (Y_l^{N_l} - Y_l)\Big\}^2\Big] = \mathbb{E}\Big[\sum_{l=1}^N (Y_l^{N_l} - Y_l)^2\Big] + 2\sum_{1\le l<q\le L} \mathbb{E}\big[(Y_l^{N_l} - Y_l)(Y_q^{N_q} - Y_q)\big],$$

and the cross terms are

$$\begin{aligned} \text{(a)} \quad \sum_{1\le l<q\le L} \mathbb{E}\big[(Y_l^{N_l} - Y_l)(Y_q^{N_q} - Y_q)\big] &= \sum_{1\le l<q\le L} \mathbb{E}(T_l^1 T_q^1) \\ \text{(b)} \quad &+ \sum_{1\le l<q\le L} \mathbb{E}(T_l^1 T_q^2) + \mathbb{E}(T_l^1 T_q^3) \\ \text{(c)} \quad &+ \sum_{1\le l<q\le L} \mathbb{E}(T_l^2 T_q^1) + \mathbb{E}(T_l^3 T_q^1) \\ \text{(d)} \quad &+ \sum_{1\le l<q\le L} \mathbb{E}(T_l^2 T_q^2) + \mathbb{E}(T_l^2 T_q^3) + \mathbb{E}(T_l^3 T_q^2) + \mathbb{E}(T_l^3 T_q^3). \end{aligned}$$

There are five new terms with respect to [4] (all those including T^3), i.e., 1 in (b), 1 in (c), and 3 in (d), but they can be dealt with similarly. In fact, since $\|\tilde{\varphi}_n\|_\infty \leq \|\overline{\varphi_n}\|_\infty \leq C\|G_n - 1\|_\infty$, and $\max\{\|G_n - 1\|_\infty^2, \|\rho_n - \rho_{n-1}\|_\infty^2\} = V_n$, the terms are all of the same type as in [4], grouped by category (a, b, c, d), and are bounded exactly as in the appendix of that paper. ∎

Appendix B. Theory related to verification of assumptions.

B.1. Restriction of prior measure. The examples in section 5 consider Gaussian prior measures μ_0, which are hence supported on an unbounded space in principle. The *restricted prior measure* is

$$\mu_{0,R}(du) := \mathbf{1}_{S_R}(u)\frac{1}{\mu_0(S_R)}\mu_0(du) , \quad S_R := \{u \in E; |u|_{L^\infty(\Omega)} \leq R\}$$

for some $R > 0$, where Ω is the spatial/temporal domain. Note that provided $\mu_0(L^\infty(\Omega)) = 1$, for any $\varepsilon > 0$, there exists a $R(\varepsilon)$ such that $|\mu_{0,R} - \mu_0|_{\mathrm{TV}} < \varepsilon$, as shown in [31]. This restriction allows for a simple verification of assumptions (A1) and (A3). In full generality one would have to carry out several technical proofs that would obscure the main ideas of the ML approach. It will be shown below that the restriction to S_R will allow (A1) and (A3) to hold for the examples considered. Note that the bound on TV-norm implies a similar bound for the difference in expectation of bounded functionals and functions with bounded second moments (via Hellinger metric, where the bound is replaced by $\varepsilon^{1/2}$, as shown in Lemma 1.30 of [26]).

Before continuing, assumption (A2) in Appendix A needs to be considered. Theorem 20 of [31] shows that under conditions on the target distribution, the Metropolis–Hastings algorithm with proposal (3.4) restricted on S_R has an $L^2(\mu_0)$ spectral gap. (See also Corollary 4 of [31] to verify that (3.4) for $\bar{u} = 0$ takes the appropriate so-called generalized pCN form.) Therefore the proposal kernel K_ℓ, conditionally on the current population of samples, satisfies a spectral gap assumption. It is beyond the scope of the present work to theoretically verify the validity of the algorithm for this weaker property (relative to (A2)), so we shall content ourselves with the stronger assumption (A2) and leave open the much more challenging question of the algorithm's rigorous validity under weaker assumptions. See also the recent work [13] for consideration of weaker assumptions in the case of the original MLSMC sampler algorithm on spaces of fixed dimensions of [4].

B.2. Conditioned diffusions. Recall (5.1). For any $T > 0$ there is this equation that has a unique solution $p \in C(\Omega, \mathbb{R}^s)$ with $\Omega = [0, T]$ and a map $u \mapsto p$ which is continuous from $C(\Omega, \mathbb{R}^s)$ to $C(\Omega, \mathbb{R}^s)$ with probability 1 under the Wiener measure. This is shown in Theorem 3.14 of [20], along with the well-posedness of the corresponding smoothing problem below. Note that since $C(\Omega, \mathbb{R}^s) \subset L^\infty(\Omega, \mathbb{R}^s)$ the prior Wiener measure can be restricted to some S_R with arbitrarily small effect.

Likewise, the path $p_{0:\ell}$ arising from the Euler–Maruyama discretization of (5.1) using the Brownian motion positions $u_{0:\ell}$ is a continuous function of $u_{0:\ell}$. The likelihood function at level ℓ will now be $\mathcal{L}_\ell(u_{0:\ell}) = \exp\left(-\frac{1}{2}|y - \mathcal{G}_\ell(u_{0:\ell})|_\Xi^2\right)$ with $\mathcal{G}_\ell(u_{0:\ell})$ denoting the mapping from the Euler scheme points $u_{0:\ell}$ to the position of $p_{0:\ell}$ at observation times. We immediately have

that $|\mathcal{G}_\ell(u_{0:\ell})| \leq C(R)$, so assumption (A1) is satisfied. We also have

$$\left|\frac{\mathcal{L}_\ell(u_{0:\ell})}{\mathcal{L}_{\ell-1}(u_{0:\ell-1})} - 1\right| \leq C(R)\left|\mathcal{G}_\ell(u_{0:\ell}) - \mathcal{G}_{\ell-1}(u_{0:\ell-1})\right| \leq C'(R) \sup_{t\in[0,T]} |p_\ell(t) - p_{\ell-1}(t)| = o(h_{\ell-1}^{\beta/2}),$$

where $p_\ell(t) = p_{\ell,i}$ for $t \in [(i-1)h_\ell, ih_\ell)$, the latter bound holding almost surely for any $\beta \in (0,1)$, as shown in Theorem 7.12 of [18]. Note that this does not provide our required uniformity in $u_{0:\ell}$ for assumption (A3); however, the required rate will be verified numerically.

B.3. Elliptic PDE inverse problem. Notice that in the case $R \to \infty$, $\mathcal{L}_\ell(u_{0:\ell})$ is not *uniformly bounded* for the full unrestricted support of the Gaussian measure μ_0. Choosing $R < \infty$, the weak form of (5.3) is continuous and coercive uniformly in u, and the Lax–Milgram lemma holds [7]. This provides the uniform bound in (A1). Uniform bounds on the PDE finite-element approximations with piecewise bilinear nodal basis functions are readily available in this case for any fixed space E_ℓ. See [4, 7, 34] for details.

Now, we proceed to extend the proof of convergence rate from finite uniform u [4] to infinite (truncated) Gaussian u. We define the V-norm as

$$|p|_V^2 := \int_{[0,1]^2} |\nabla p(u)|^2 dx, \quad p \in V,$$

noting that the boundary condition (5.4) guarantees that $\int_\Omega p dx = 0$ and so Poincaré inequality applies. As in [4], the quantity we would like to bound uniformly in u is

$$|p_\ell(u_{0:\ell}) - p(u)|_V \leq |p_\ell(u_{0:\ell}) - p(u_{0:\ell})|_V + |p(u_{0:\ell}) - p(u)|_V. \tag{B.1}$$

The first term is dealt with as in [4]. The second term comes from the truncation to E_ℓ. Denote $\bar{p} = p(u_{0:\ell})$ and observe that for all $v \in V$

$$\langle \nabla v, \mathbb{K}_\ell(u_{0:\ell})\nabla\bar{p} - \mathbb{K}(u)\nabla p\rangle = 0,$$

so

$$\langle \nabla v, \mathbb{K}_\ell(u_{0:\ell})\nabla\bar{p} - \mathbb{K}_\ell(u_{0:\ell})\nabla p\rangle + \langle \nabla v, \mathbb{K}_\ell(u_{0:\ell})\nabla p - \mathbb{K}(u)\nabla p\rangle = 0.$$

Letting $v = \bar{p} - p$ and rearranging, we have (using also Poincaré inequality)

$$\begin{aligned} |\bar{p} - p|_V^2 &\leq C(|u_{0:\ell}|_\infty)|(\mathbb{K}_\ell(u_{0:\ell}) - \mathbb{K}(u))(\nabla p)\cdot(\nabla(\bar{p} - p))| \\ &\leq C(|u_{0:\ell}|_\infty)|\mathbb{K}_\ell(u_{0:\ell}) - \mathbb{K}(u)|_{L^\infty(\Omega)}|p|_V|\bar{p} - p|_V. \end{aligned}$$

Therefore on the truncated space S_R, the following holds:

$$|\bar{p} - p|_V \leq C(R)|\mathbb{K}_\ell(u_{0:\ell}) - \mathbb{K}(u)|_{L^\infty(\Omega)} = \mathcal{O}(h_\ell^{\beta/2}) \tag{B.2}$$

for some $\beta \in (2,4)$. (See section 3.3 of [7].) The error due to the solution of the PDE with finite element discretization of diameter h_ℓ is also given by

$$|p_\ell(u_{0:\ell}) - p(u_{0:\ell})|_V = \mathcal{O}(h_\ell^{\beta/2})$$

for $\beta \in (2,4)$ [4, 7]. Ultimately, the quantity

$$V_\ell = \max\{\|G_\ell - 1\|_\infty^2, \|\rho_\ell - \rho_{\ell-1}\|_\infty^2\}$$

can be bounded by Ch_ℓ^β, as both terms are controlled by (B.1). The first term is handled similarly to the work [4]. Typically the functions ρ_ℓ we are interested in will have the form $\rho_\ell(u_{0:\ell})_i = f_i(p_\ell(u_{0:\ell}))$ for some $f_i \in V^*$, and hence $V_\ell = \mathcal{O}(\|\|p_\ell(u_{0:\ell}) - p(u)|_V\|_\infty)$.

REFERENCES

[1] A. Beskos, D. Crisan, and A. Jasra (2014), *On the stability of sequential Monte Carlo methods in high dimensions*, Ann. Appl. Probab., 24, pp. 1396–1445.

[2] A. Beskos, A. Jasra, E. Muzaffer, and A. Stuart (2015), *Sequential Monte Carlo methods for Bayesian elliptic inverse problems*, Statist. Comput., 25, pp. 727–737.

[3] A. Beskos, M. Girolami, S. Lan, P Farrell, and A. Stuart (2017), *Geometric MCMC for infinite-dimensional inverse problems*, J. Comput. Phys., 335, pp. 327–351.

[4] A. Beskos, A. Jasra, K. J. H. Law, R. Tempone, and Y. Zhou (2017), *Multilevel sequential Monte Carlo samplers*, Stochastic Process. Appl., 127, pp. 1417–1440.

[5] C. Bierig and A. Chernov (2015), *Convergence analysis of multilevel Monte Carlo variance estimators and application for random obstacle problems*, Numer. Math., 130, pp. 579–613.

[6] Y. Chen, D. Keyes, K. J. H. Law, and H. Ltaief (2016), *Accelerated dimension-independent adaptive Metropolis*, SIAM J. Sci. Comput., 38, pp. 539–565.

[7] P. G. Ciarlet (2002), *Finite Element Method for Elliptic Problems*, Classics Appl. Math., SIAM, Philadelphia.

[8] S. Cotter, G. O. Roberts, A. M. Stuart, and D. White (2014), *MCMC methods for functions: Modifying old algorithms to make them faster*, Statist. Sci., 28, pp. 424–446.

[9] T. Cui, K. J. Law, and Y. M. Marzouk (2016), *Dimension-independent likelihood-informed MCMC*, J. Comput. Phys., 304, pp. 109–137.

[10] T. Cui, J. Martin, Y. M. Marzouk, A. Solonen, and A. Spantini (2014), *Likelihood-informed dimension reduction for nonlinear inverse problems*, Inverse Problems, 11, 114015.

[11] P. Del Moral (2004), *Feynman–Kac Formulae: Genealogical and Interacting Particle Systems with Applications*, Springer, New York.

[12] P. Del Moral, A. Doucet, and A. Jasra (2006), *Sequential Monte Carlo samplers*, J. Roy. Statist. Soc. Ser. B, 68, pp. 411–436.

[13] P. Del Moral, A. Jasra, and K. J. H. Law (2017), *Multilevel sequential Monte Carlo: Mean square error bounds under verifiable conditions*, Stoch. Anal. Appl., 35, pp. 478–498.

[14] P. Del Moral, A. Jasra, K. J. H. Law, and Y. Zhou (2017), *Multilevel sequential Monte Carlo samplers for normalizing constants*, ACM Trans. Model. Comput. Simul., 27, 20.

[15] D. Donoho, M. Gavish, and I. Johnstone (2018), *Optimal shrinkage of eigenvalues in the spiked covariance model*, Ann. Statist. to appear.

[16] W. Förstner and B. Moonen (2003), *A metric for covariance matrices*, Geodesy—The Challenge of the 3rd Millennium, Springer, New York, pp. 299–309.

[17] M. B. Giles, (2008), *Multilevel Monte Carlo path simulation*, Oper. Res., 56, pp. 607–617.

[18] C. Graham and D. Talay (2013), *Stochastic Simulation and Monte Carlo Methods: Mathematical Foundations of Stochastic Simulation*, Stoch. Model. Appl. Probab. 68, Springer, New York.

[19] A. Gregory, C. J. Cotter, and S Reich (2016), *Multilevel ensemble transform particle filtering*, SIAM J. Sci. Comput., 38, A1317–A1338.

[20] M. Hairer, A. M. Stuart, and J. Voss (2011), *Signal processing problems on function space: Bayesian formulation, stochastic PDEs and effective MCMC methods*, in The Oxford Handbook of Nonlinear Filtering, Oxford University Press, Oxford, pp. 833–873.
[21] S. Heinrich (1998), *Monte Carlo complexity of parametric integration*, J. Complexity, 14, pp. 151–175.
[22] H. Hoel, K. J. Law, and R. Tempone (2016), *Multilevel ensemble Kalman filtering*, SIAM J. Numer. Anal., 54, pp. 1813–1839.
[23] V. H. Hoang, C. Schwab, and A. M. Stuart (2013), *Complexity analysis of accelerated MCMC methods for Bayesian inversion*, Inverse Problems, 29, 085010.
[24] A. Jasra, K. Law, and Y. Zhou (2016), *Forward and inverse uncertainty quantification using multilevel Monte Carlo algorithms for an elliptic nonlocal equation*, Int. J. Uncertain. Quantif., 6, pp. 501–514.
[25] A. Jasra, D. A. Stephens, A. Doucet, and T. Tsagaris (2011), *Inference for Lévy driven stochastic volatility models via adaptive sequential Monte Carlo*, Scand. J. Stat., 38, pp. 1–22.
[26] K. Law, A. Stuart, and K. Zygalakis (2015), *Data Assimilation*, Texts Appl. Math., Springer, New York.
[27] F. Li, Q. Dai, Q. Xu, and G. Er (2009), *Weighted subspace distance and its applications to object recognition and retrieval with image sets*, IEEE Signal Process. Lett., 16, pp. 227–230.
[28] Y. Marzouk and O. Zahm, *private communications*, 2016.
[29] K. L. Mengersen and R. L. Tweedie (1996), *Rates of convergence of the Hastings and Metropolis algorithms*, Ann. Statist., 24, pp. 101–121.
[30] C. H. Rhee and P. W. Glynn (2015), *Unbiased estimation with square root convergence for SDE models*, Oper. Res., 63, pp. 1026–1043.
[31] D. Rudolf and B. Sprungk (2018), *On a generalization of the preconditioned Crank–Nicolson Metropolis algorithm*, Found. Comput. Math., 18, pp. 309–343.
[32] A. Spantini, A. Solonen, T. Cui, J. Martin, T. Tenorio, and Y. Marzouk (2015), *Optimal low-rank approximations of Bayesian linear inverse problems*, SIAM J. Sci. Comput., 37, A2451–A2487.
[33] A. M. Stuart (2010), *Inverse problems: A Bayesian perspective*, Acta Numer., 19, pp. 451–559.
[34] O. C. Zienkiewicz (1977), *The Finite Element Method*, McGraw-Hill, London.